Planning and interpreting the sample size of trials with multiple outcomes

Papageorgiou, Spyridon N

Abstract: An orthodontist participates in an international orthodontic conference and hears from the company representatives of a new experimental custom-made appliance for fixed appliance treatment. According to its manufacturer, using this experimental appliance would bring multiple benefits including less discomfort, better oral hygiene, and reduced treatment duration compared to conventional appliances.

DOI: https://doi.org/10.1177/1465312519831196

STATISTICAL CORNER

Planning and interpreting the sample size of trials with multiple outcomes

Spyridon N. Papageorgiou

Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

ORCID Spyridon N. Papageorgiou http://orcid.org/0000-0003-1968-3326

CONTACT Spyridon N. Papageorgiou Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich CH 8032, Switzerland; snpapage@gmail.com.

Words in text: 1382

Disclosure statement

No potential conflict of interest was reported by the author.

Theoretical scenario

An orthodontist participates in an international orthodontic conference and hears from the company representatives of a new experimental custom-made appliance for fixed appliance treatment. According to its manufacturer, using this experimental appliance would bring multiple benefits including less discomfort, better oral hygiene, and reduced treatment duration compared to conventional appliances.

Intrigued by these extraordinary claims, the orthodontist builds a multicentre research team and secures a grant from the local orthodontic scientific society to perform a large clinical trial that might provide definite evidence to support or reject these claims. A randomized clinical trial is set up with two parallel patient groups: one group will receive the experimental appliances and the other the conventional (control) appliances. At the planning phase, the statistician involved in the research team suggests to the clinicians that the number of patients that need to be recruited in the trial needs to be defined a priori before the trial's start. This will give the trial enough statistical power to clinically identify a difference between the performance of the two appliances, if such a difference actually exists.

The orthodontists involved in the research team agree that the overall treatment duration from insertion to removal of the fixed appliances is the most relevant endpoint for both patient and doctor, and should therefore be the trial's primary outcome. Based on data from an existing trial with similar clinical setting to the planned trial (Yassir et al. 2018), the average expected fixed appliance treatment duration was found to be 29.3 months with a Standard Deviation (SD) of 9.5 months. Using their clinical subjective judgement, the researchers argue that the experimental appliances cost considerably more than the control appliances and should therefore reduce treatment duration by at least 25% (7.3 months in our case) to justify these increased costs. Adopting a two-sided type I error of 5% and type II error of 20% (i.e. a statistical power of 80%), the statistician performed a sample size calculation for a Student's test for 2 independent groups and found that a total of 56 patients (28 patients in each group) would need to be recruited in the trial (Table 1).

Table 1. Sample size calculation based for the theoretical trial's primary outcome (overall treatment duration) with alpha set at 5% and beta at 20%, based on a two-sided p value from a Student's t-test for two independent samples. Needed sample size is inflated by 15% (and rounded up to the next even number) to account for possible dropouts.

Baseline data					**Expectation**			**Sample needed**	
Source	**Outcome**	**Mean**	**SD**		**Expected benefit**	**Expected mean**		**For analysis**	**Expecting dropouts**

Yassir et al. 2018	Total treatment duration (months)	29.30	9.50		-25%	21.98		56	66

SD, standard deviation.

After consulting with the statistician, the orthodontists set an aim to recruit 66 (33 patients in each group) patients in total to account for an expected 15% dropout rate. Wanting to make the most out of the existing material, they add in the trial's protocol the following secondary outcomes:

(i) patient discomfort during alignment / levelling: measured with a Likert-scale questionnaire 24 hours after insertion of each aligning archwire;

(ii) gingival health: measured with Löe's index 1 year after appliance insertion;

(iii) duration of the alignment / levelling phase: measured in months from appliance insertion up to insertion of the working archwire

(iv) duration of the working / finishing phase: measured in months from insertion of the working archwire to removal of the appliances.

The orthodontists initiate the trial and recruit a total of 66 patients, from which 10 patients either drop out or are excluded from the final analysis due to missing data. At the end, data from 56 patients (28 in each group) are analysed statistically and indicate that no statistically significant difference ($P>0.05$) can be found between the two appliances for any of the trial's primary or secondary outcomes.

Which of the following statements are true, if any?

(A) A sample size calculation was performed, so the present trial was probably adequately powered to identify a potential 25% benefit of the experimental appliance for the primary outcome of overall treatment duration, if such exists.

(B) A sample size calculation was performed, so the present trial was probably adequately powered to identify a potential 25% benefit of the experimental appliance for any of the trial's secondary outcomes, if such exists.

(C) As no significant difference was seen between the two groups of the present trial, we can be fully confident that use of the experimental appliance is not associated with any benefits for any of the primary or secondary outcomes.

Discussion

The authors of the current trial performed an a priori sample size calculation that was based on data from a trial with clinical setting similar to the setting of their own trial, adopted conventional approaches for type I and type II errors, and made what seemed to them as reasonable clinically relevant assumptions on the expected treatment effect. They were prudent enough to predict that patient dropouts were to be expected in a follow-up of more than two years and ended up analysing at trial's end the exact number of patients they had set as goal beforehand. Therefore, it is reasonable to assume that the present trial was adequately powered to detect a difference of 25% (7.3 months) in overall treatment duration, if such an effect exists. So, statement (A) is true.

However, this does not mean that the current trial is de factor adequately powered to investigate existing differences in any outcome. The authors of the current trial performed an a priori sample size calculation only for the trial's primary outcome. Had the authors informed their statistician that another four secondary outcomes would be assessed in their trial and that the trial should be adequately powered for all of them, then additional sample size calculations would be needed. Based on previous trials with similar clinical settings (Ong et al. 2011; Kaklamanos et al. 2017; Yassir et al. 2018), under the same assumptions as for the primary outcome, and after accounting for dropouts the authors would need to recruit somewhere between 28 patients and 162 patients in total according to each outcome (Table 2). If all outcomes were to be adequately powered, then the authors would need to meet the highest requirement—that is, they would need to recruit 162 patients overall. If the assumptions made for the current sample size calculations hold true, then the finally analysed sample of 56 patients in the current trial is adequately powered only for the primary outcome of treatment duration and the secondary outcome of gingival health at 1 year. It is however, somewhat underpowered to identify a benefit in terms of patient discomfort (60 patients needed) and severely underpowered to identify a benefit in terms of alignment duration or working / finishing duration (94 or 140 patients needed). This means that even if a substantial difference of at least 25% for any of these secondary outcomes actually exists, the present trial might not be able to find it or might find results that deviate from the truth, as seen previously (Papageorgiou, 2018). Therefore, statement (B) is (at least partly) false, as the trial is underpowered for three of the secondary outcomes.

Table 2. Sample size calculation based for the theoretical trial's primary outcome (overall treatment duration) and all secondary outcomes (discomfort, gingival health, and duration of specific treatment

phases) with alpha set at 5% and beta at 20%, based on a two-sided p value from a Student's t-test for two independent samples. Needed sample size is inflated by 15% (and rounded up to the next even number) to account for possible dropouts.

Baseline data					Assumption			Sample needed	
Source	**Outcome**	**Mean**	**SD**		**Assumed benefit**	**Assumed mean**		**Total for 2 groups**	**Expecting dropouts**
Yassir et al. 2018	Total treatment duration (months)	29.30	9.50		-25%	21.98		56	66
Ong et al. 2011	Discomfort at 24h (Likert scale)	3.30	1.10		-25%	2.48		60	70
Kaklamanos et al. 2017	Gingival index 1 year in treatment (Löe's index)	1.53	0.31		-25%	1.15		24	28
Yassir et al. 2018	Duration of alignment / levelling phase (months)	11.8	5.0		-25%	8.85		94	110
Yassir et al. 2018	Duration of working / finishing phase (months)	17.40	9.10		-25%	13.05		140	162

SD, standard deviation

Sample size calculations in clinical trials are considered as important by funding agencies, ethics review boards, and journals. This is understandable, since underpowered trials might be considered unethical by many people, as patients might be subjected to unnecessary risks by taking part in studies that cannot realise their full scope or might provide misleading results. At the same time, it must be stressed out that sample size calculations in clinical trials might be plagued by inaccurate assumptions of baseline risk, erroneous calculations, a posteriori set hypothesised treatment effects, and incomplete reporting (Charles et al. 2009; Koletsi et al. 2014). Even if correctly executed, sample size calculation is based on statistical and clinical assumptions that might be questionable (Guyatt et al. 2008), while issues of feasibility usually play a major role when calculating the needed sample for a clinical trial. In the specific theoretical example, the current trial gives some evidence that the experimental appliance might not be associated with a benefit of the hypothesised magnitude in overall duration or gingival health. We are even less assured by provided evidence about the true effect of the experimental appliance as far as the patient discomfort, alignment phase duration, or working / finishing phase duration is concerned. Also, any single trial – regardless of how well powered it is – is unlikely to be able to definitively answer a clinical question and synthesis of multiple clinical trials with low risk of bias remains the gold standard. So, statement (C) is broadly seen false.

Finally, it is worth of note that sample size calculations are very sensitive on the assumptions made, especially about the expected treatment effects, as will be illustrated in the next Statistical Corner.

References

Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. 2009. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 338:b1732.

Guyatt GH, Mills EJ, Elbourne D. 2008. In the era of systematic reviews, does the size of an individual trial still matter. PLoS Med. 5(1):e4.

Kaklamanos EG, Mavreas D, Tsalikis L, Karagiannis V, Athanasiou AE. 2017. Treatment duration and gingival inflammation in Angle's Class I malocclusion patients treated with the conventional straight-wire method and the Damon technique: a single-centre, randomised clinical trial. J Orthod. 44(2):75—81.

Koletsi D, Fleming PS, Seehra J, Bagos PG, Pandis N. 2014. Are sample sizes clear and justified in RCTs published in dental journals? PLoS One. 9(1):e85949.

Ong E, Ho C, Miles P. 2011. Alignment efficiency and discomfort of three orthodontic archwire sequences: a randomized clinical trial. J Orthod. 38(1):32—39.

Papageorgiou SN. 2018. On the sample size of clinical trials – revisited. J Orthod. 45(4):296—298.

Yassir YA, El-Angbawi AM, McIntyre GT, Revie GF, Bearn DR. 2018. A randomized clinical trial of the effectiveness of 0.018-inch and 0.022-inch slot orthodontic bracket systems: part 1-duration of treatment. Eur J Orthod. doi: 10.1093/ejo/cjy037. [Epub ahead of print]